PATENT ABSTRACTS OF JAPAN

(11)Publication number : 08-328759
(43)Date of publication of application : 13.12.1996

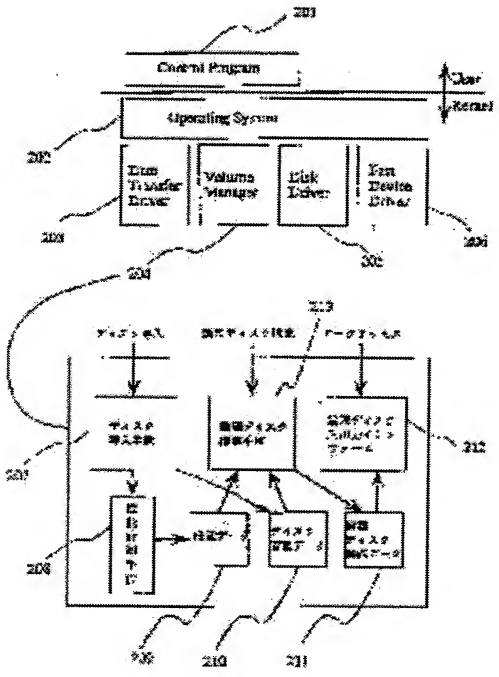
(51)Int.CI. G06F 3/06
G06F 3/06
G06F 13/10

(21)Application number : 07-130555
(22)Date of filing : 29.05.1995

(71)Applicant : MITSUBISHI ELECTRIC CORP
(72)Inventor : SAKAKURA TAKASHI
FUSHIMI SHINYA

(54) INPUT/OUTPUT PROCESSING SYSTEM

(57)Abstract:
PURPOSE: To improve the throughput of the entire system by generating logical disk management data for which a stripe width for uniformizing response time required for the input/output of data for one stripe of respective disk devices is set.
CONSTITUTION: The performance characteristics of the respective disk devices for constituting the stripe are recognized by a performance measurement means 208. It is performed by using a parameter, making the disk devices execute input/output operations and measuring the response time and the result is recorded. Then, based on the measured performance data 209 of the respective disk devices, a data amount to be arranged to the respective disk devices so as to uniformize each response time required for the input/output of the data for one stripe of the disk devices for constituting a disk drive, that is the stripe width, is decided. Then, logical disk constitution data 211 according to it are constituted by a logical disk constitution means 213, that is, the logical disk device of striping constitution is constituted.



LEGAL STATUS

[Date of request for examination]	19.11.1999
[Date of sending the examiner's decision of rejection]	13.02.2001
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]	
[Date of final disposal for application]	
[Patent number]	3201219
[Date of registration]	22.06.2001
[Number of appeal against examiner's decision of rejection]	2001-03985

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平8－328759

(43)公開日　平成８年(1996)12月13日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ	3/06	５４０		Ｇ０６Ｆ	3/06	５４０
		３０１				３０１Ｍ
	13/10	３４０	7368－5Ｅ		13/10	３４０Ｂ

審査請求　未請求　請求項の数17　ＯＬ　（全 22 頁）

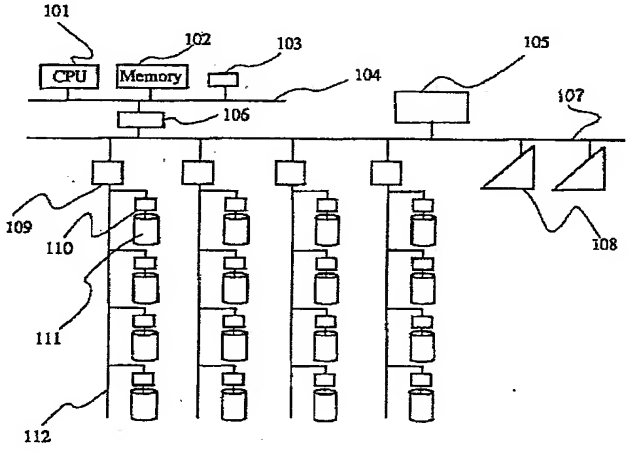
(21)出願番号　特願平7－130555

(22)出願日　平成７年(1995) ５月29日

(71)出願人　000006013
三菱電機株式会社
東京都千代田区丸の内二丁目２番３号
(72)発明者　坂倉　隆史
鎌倉市大船五丁目１番１号　三菱電機株式会社情報システム研究所内
(72)発明者　伏見　信也
鎌倉市大船五丁目１番１号　三菱電機株式会社情報システム研究所内
(74)代理人　弁理士　宮田　金雄　（外３名）

(54)【発明の名称】　入出力処理システム

(57)【要約】
【目的】　通常のディスク装置を利用し高速データ転送サブシステムを実現する。
【構成】　ソフトウェア制御によるディスク間同期コストを低減したストライプトディスクとタイムアウトによる同期制御手段、さらに、ユーザメモリを使用しないデータ転送手段により高速大量データ転送システムを実現する。



1

【 特許請求の範囲】

【 請求項1 】　入出力システムを構成する複数のディスク装置の性能特性をシステム管理者から与えられる性能データもしくは前記ディスク装置を動作させて性能を計測する性能計測手段により収集する性能データ収集手段と、この性能データ収集手段で収集した性能データを基に前記複数のディスク装置を用いて論理ディスク装置を構成する論理ディスク構成手段と、を有する論理ディスク制御手段を備え、

前記論理ディスク構成手段は、前記論理ディスク装置を構成する各ディスク装置の1ストライプ分のデータ入出力に要する応答時間が均等になるストライプ幅を設定した論理ディスク管理データを生成し、前記論理ディスク制御手段は、前記論理ディスク管理データにより前記論理ディスク装置を制御することを特徴とする入出力処理システム。

【 請求項2 】　入出力装置に対する入出力命令発行時に、前記入出力装置の動作に要する制限時間を設定する制限時間設定手段と、設定した制限時間が経過したら前記入出力命令により起動された入出力動作を終了させる手段と、を備えたことを特徴とする入出力処理システム。

【 請求項3 】　前記制限時間設定手段で設定される時間を設定するタイマと、このタイマからの制限時間終了を受け入出力動作を終了させる手段と、を入出力装置の制御をする制御装置に設けたことを特徴とする請求項2 に記載の入出力処理システム。

【 請求項4 】　前記入出力装置は複数のディスク装置から構成された論理ディスク装置であることを特徴とする請求項2 または請求項3 に記載の入出力処理システム。

【 請求項5 】　請求項1 に記載の入出力システムにおいて、入出力装置に対する入出力命令発行時に、この入出力命令の処理に関する制限時間を設定する制限時間設定手段と、設定した制限時間が経過したら前記入出力命令により起動された入出力動作を終了させる処理終了手段と、を設けたことを特徴とする入出力処理システム。

【 請求項6 】　前記性能計測手段は、各ディスク装置に対する入出力命令実行時の応答時間が、最短、あるいは、最長となる条件を設定して、入出力命令を実行し、その応答時間を計測することを特徴とする請求項1 または請求項5 に記載の入出力処理システム。

【 請求項7 】　前記性能計測手段は、システムへのディスク装置追加時にこのディスク装置に対する初期化処理の一部としてこのディスク装置の性能測定を行うことを特徴とする請求項1 または請求項5 または請求項6 に記載の入出力処理システム。

【 請求項8 】　前記性能収集手段は、複数のディスク装置が接続された入出力バスの構成、バス転送性能をシステム管理者より与えられたシステム構成データ、あるいは、前記性能計測手段により前記バスに接続された入出力装置を実際に動作させることによってバス性能データを収集し、前記論理ディスク構成手段が、この収集した性能データを基に、各々のディスク性能と接続されたバス転送性能を考慮して、論理ディスク装置を構成することを特徴とする請求項1 または請求項5 乃至請求項7 のいずれかに記載の入出力処理システム。

2

【 請求項9 】　前記論理ディスク構成手段は、論理ディスク装置を構成するディスク装置に割り当てられる1回あたりの入出力命令におけるデータ転送量をそのディスク装置の1トラックの中に収めることを特徴とする請求項1 または請求項5 乃至請求項9 のいずれかに記載の入出力処理システム。

【 請求項1０】　前記論理ディスク構成手段は、外内周部で性能差のある均質な複数のディスク装置により論理ディスクを構成する時は、各々のディスク装置の外内周部を交互に組合せて構成することにより論理ディスクの見かけの入出力性能がデータの配置されたディスク装置上の位置によらず均一となることを特徴とする請求項9 に記載の入出力処理システム。

【 請求項1１】　前記論理ディスク制御手段は、論理ディスク装置に対する入出力要求があった場合、その論理ディスク装置を構成する全てのディスク装置への入出力命令発行直前のそのディスク装置のヘッド位置等の状態を動的に判断して、その論理ディスク装置への入出力命令実行時間が最短となるように入出力命令を発行することを特徴とする請求項1 または請求項5 乃至請求項1０のいずれかに記載の入出力処理システム。

【 請求項1２】　前記論理ディスク制御手段は、複数のディスク装置の各々に複数回の入出力要求が発生する入出力サイズで論理ディスク装置に対して入出力要求があった時は、この論理ディスク装置を構成するディスク装置の性能特性、および、ディスクが接続された入出力バスの特性を判断し、必要に応じて、同一ディスク装置への複数入出力要求をより少ない入出力要求回数とする、あるいは、該複数ディスク装置への入出力要求中に同期ポイントを設けることにより、該論理ディスク装置への入出力時間が最短となるように入出力命令を発行することを特徴とする請求項1１に記載の入出力処理システム。

【 請求項1３】　前記論理ディスク制御手段は、システム管理者より与えられた性能を満たす論理ディスク装置を自動的に構成することを特徴とする請求項1 または請求項5 乃至請求項1２のいずれかに記載の入出力処理システム。

【 請求項1４】　入出力装置に対するデータ転送制御を行うデータ転送ドライバを備え、このデータ転送ドライバは入出力命令の入出力対象装置に論理ディスク装置が含まれている場合は、システムメモリ上にデータ転送用の2面バッファを確保してデータ転送を行うことを特徴とする請求項1 乃至請求項1３のいずれかに記載の入出

力処理システム。

【 請求項１５】　入出力装置の接続される入出力バスの制御をするバスアービタを設け、このバスアービタはデータ転送元及びデータ転送先の装置のＩＤを登録するソースレジスタ及びデスティニレジスタを備え、前記データ転送ドライバは、入出力命令の入出力対象装置に論理ディスク装置が含まれていない場合は、前記バスアービタを駆動させることによりシステムメモリを利用せずにデータ転送を行うことを特徴とする請求項１４に記載の入出力処理システム。

【 請求項１６】　前記データ転送ドライバは、入出力命令で指定された入力装置と出力装置のデータ転送速度に差がある場合は、システムメモリ上にデータ転送用のバッファを確保して前記バスアービタを駆動することを特徴とする請求項１４に記載の入出力処理システム。

【 請求項１７】　前記論理ディスク装置を構成する複数のディスク装置が各々異なる入出力バスに接続されて複数の論理ディスク装置を構成する請求項1 または請求項５乃至請求項１３のいずれかに記載の入出力処理システムにおいて、前記各入出力バスに接続された各々のディスク装置を制御するディスク制御装置に同じ入出力バスに接続されたディスク装置間での複写手段を設け、異なる論理ディスク装置間でデータの複写を行うことを特徴とする入出力処理システム。

【 発明の詳細な説明】

【０００１】

【 産業上の利用分野】本発明は計算機システムにおける入出力処理システムに関し、特に大量データを高速に入出力するための入出力処理システムに関する。

【０００２】

【 従来の技術】近年、磁気ディスク記憶装置の小型化、低価格化に伴い、ＲＡＩＤと呼ばれるディスクサブシステムがパーソナルコンピュータやワークステーションに普及してきた。ＲＡＩＤは複数のディスク装置で論理的なディスク装置を構成して性能や信頼性の向上を図るもので、使用目的によって、一般的にＲＡＩＤ０レベルからＲＡＩＤ５レベルに区分される構成方式のいずれかが採用される。ＲＡＩＤは一般的にディスクコントローラの機能として実現され、パーソナルコンピュータではＳＣＳＩホストアダプタの機能として実現されることが多い。

【０００３】一方、ハードウェアによらずソフトウェアによる制御で、複数のディスク装置により論理的なディスク装置を構成する技術も知られている。ＩＢＭ社のロジカルボリュームマネージャ、ベリタス社のボリュームマネージャなどが、その代表的な製品であり、これらにおいては、ディスクコンカティネーティング、ディスクミラーリング、ディスクストライピング等の機能が提供されている。また、製品によってはＲＡＩＤ機能を全てカバーすべく試みているものもある。

【０００４】ＲＡＩＤシステムにおいては、大量で高速な入出力のためにはＲＡＩＤ３レベルあるいはＲＡＩＤ０レベルが用いられる。ＲＡＩＤ３レベルはバイトインターリーブ構成でパリティデータを付加したデータレイアウト構成をとる。ＲＡＩＤ０レベルはソフトウェアによるディスクストライピングと同等でパリティデータが付加されない。

【０００５】ディスクストライピングやＲＡＩＤにより期待できる入出力性能上の効果は、インターリーブ配置されたデータに対して複数のディスク装置を含むようなデータサイズの入出力要求があった時に、各々の当該ディスク装置に入出力要求を振り分け、ディスク装置を並列動作させることによってもたらされるユーザから見た入出力性能の向上と、論理ディスク装置に対するコンカレントアクセスを複数のディスク装置に振り分けることによってもたらされるシステムとしてのスループット性能の向上があるが、大量データの高速入出力のためには前者の効果を追求することになる。

【０００６】ところで、ストライプ構成された論理ディスク装置に対する入出力は、それぞれの当該ディスク装置に振り分けられるが、１回の論理ディスク装置に対する入出力ごとに振り分けられた全ての入出力に対して同期をとらねばならない。並列に動作するディスク装置の処理時間にばらつきがあると、処理終了の待ち合わせ時間をそのばらつきの分だけ必要になり、論理ディスク装置の入出力性能の劣化につながる。この同期を乱す要因としては、個々のディスク装置そのものの性能差、シーク（位置決め）、回転待ち等に起因する動的な性能のばらつき、データバスの待ち合わせ等がある。

【０００７】ディスク装置間の待ち合わせのオーバーヘッドを削減するためにＲＡＩＤシステムにおいては、例えば、ＲＡＩＤ３レベルの適用されたシステムでは、均質なディスク装置でシステムを構成し、ディスクの回転待ちによる同期の乱れをなくすため、全てのディスクの回転を同期させ、各ディスク装置への入出力命令の遅延を吸収するために、データを配置するセクタ位置を少しずつずらすといった技術が知られているが、これには、一つのハードウェアサブシステムとしてＲＡＩＤを設計できるという前提が実現条件となる（ＭＩＣＨＥＬＬＥ Ｙ．ＫＩＭ，“Ｓｙｎｃｈｒｏｎｉｚｅｄ Ｄｉｓｋ Ｉｎｔｅｒｌｅａｖｉｎｇ”，ＩＥＥＥ ＴＲＡＮＳＡＣＴＩＯＮＳ ＯＮ ＣＯＭＰＵＴＥＲＳ．ＶＯＬ．Ｃ－３５，ＮＯ．１１，ＮＯＶＥＭＢＥＲ １９８６）。

【０００８】また、さらに細かい視点で、ディスク装置間の同期を乱す要因として、ディスク装置のヘッドの位置ずれ補正をとりあげ、ある基準となるディスクにおいて、位置ずれ補正が必要になった時点で、ディスクアレイを構成する全てのディスク装置に位置ずれ補正を促すコマンドを発行し、位置ずれ補正による同期の乱れを最

5

小限に押さえる技術が特開平5 －２７９１０に開示されている。

【０００９】また、特開平5 －２５７６１１号公報には、ＲＡＩＤシステムにおいて、論理ディスク装置をソフトウェア制御により、柔軟な構成で実現することが開示されている。即ち、ディスクアレイのパーティショニングを目的として、柔軟に1 つあるいは複数のディスクアレイ上にＲＡＩＤレベルの異なる論理ディスク装置を混在させ、なおかつ、ディスク装置間の同期を乱すことによる性能劣化を押さえるデータレイアウト方法についての技術が開示されている。

【００１０】

【発明が解決しようとする課題】以上のように、従来においては、論理ディスク装置をソフトウエア制御により論理ディスク装置を構成するものは存在したが、最適な論理ディスク装置を構成するために、論理ディスク装置を構成するディスク装置の性能を測定して、そのデータを基に、論理ディスク装置を構成する技術はなかった。

【００１１】本発明は、効率のよい入出力処理システムを提供し、システム全体のスループットを向上させることを目的とする。より具体的には、ＲＡＩＤを使用せずに、一般的なディスク装置を使用して最適な論理ディスク装置を構成して入出力システムの性能を向上させることを目的とする。また、画像データのような大量データの処理においては、１回の入出力データ量が非常に大きなものとなるが、この場合1 回の入出力命令に対し指定された大量のデータが完全に転送されるのを待つよりも、指定された量よりも少ないデータを受取り、より早期に処理が開始できる入出力処理システムを提供することを目的とする。さらに、入出力装置間の転送において、効率のよい転送が可能な入出力システムを提供することを目的とする。

【００１２】

【課題を解決するための手段】第1 の発明に係わる入出力処理システムは、入出力システムを構成する複数のディスク装置の性能特性をシステム管理者から与えられる性能データもしくは前記ディスク装置を動作させて性能を計測する性能計測手段により収集する性能データ収集手段と、この性能データ収集手段で収集した性能データを基に前記複数のディスク装置を用いて論理ディスク装置を構成する論理ディスク構成手段と、を有する論理ディスク制御手段を設け、前記論理ディスク構成手段は、前記論理ディスク装置を構成する各ディスク装置の1 ストライプ分のデータ入出力に要する応答時間が均等になるストライプ幅を設定した論理ディスク管理データを生成し、前記論理ディスク制御手段がこの論理ディスク管理データにより前記論理ディスク装置を制御するようにしたものである。

【００１３】第2 の発明に係わる入出力処理システムは、入出力装置に対する入出力命令発行時に、入出力装置の動作に要する制限時間を設定する制限時間設定手段と、設定した制限時間が経過したら前記入出力命令により起動された入出力動作を終了させる手段と、をもうけるようにしたものである。

6

【００１４】第3 の発明に係わる入出力処理システムは、前記制限時間設定手段で設定される時間を設定するタイマと、このタイマからの制限時間終了を受け入出力動作を終了させる手段と、を入出力装置の制御をする制御装置に設けるようにしたものである。

【００１５】第4 の発明に係わる入出力処理システムは、前記入出力装置を複数のディスク装置から構成された論理ディスク装置で構成するようにしたものである。

【００１６】第5 の発明に係わる入出力処理システムは、第1 の発明における入出力システムにおいて、入出力装置に対する入出力命令発行時に、この入出力命令の処理に関する制限時間を設定する制限時間設定手段と、設定した制限時間が経過したら前記入出力命令により起動された入出力動作を終了させる処理終了手段と、を設けようにしたものである。

【００１７】第6 の発明に係わる入出力処理システムは、前記性能計測手段が、各ディスク装置に対する入出力命令実行時の応答時間が、最短、あるいは、最長となる条件を設定して、入出力命令を実行し、その応答時間を計測するようにしたものである。

【００１８】第7 の発明に係わる入出力処理システムは、前記性能計測手段が、システムへのディスク装置追加時にこのディスク装置に対する初期化処理の一部としてこのディスク装置の性能測定を行うようにしたものである。

【００１９】第8 の発明に係わる入出力処理システムは、前記性能収集手段が、複数のディスク装置が接続された入出力バスの構成、バス転送性能をシステム管理者より与えられたシステム構成データ、あるいは、前記性能計測手段により前記バスに接続された入出力装置を実際に動作させることによってバス性能データを収集し、前記論理ディスク構成手段が、この収集した性能データを基に、各々のディスク性能と接続されたバス転送性能を考慮して、論理ディスク装置を構成するようにしたものである。

【００２０】第9 の発明に係わる入出力処理システムは、前記論理ディスク構成手段が、論理ディスク装置を構成するディスク装置に割り当てられる１回あたりの入出力命令におけるデータ転送量をそのディスク装置の1 トラックの中に収めるようにしたものである。

【００２１】第１０の発明に係わる入出力処理システムは、前記論理ディスク構成手段が、外内周部で性能差のある均質な複数のディスク装置により論理ディスクを構成する時は、各々のディスク装置の外内周部を交互に組合せて構成することにより論理ディスク装置の見かけの入出力性能がデータの配置されたディスク上の位置によ

らず均一となるようにしたものである。

【００２２】第１１の発明に係わる入出力処理システムは、前記論理ディスク制御手段が、論理ディスク装置に対する入出力要求があった場合、その論理ディスク装置を構成する全てのディスク装置への入出力命令発行直前のそのディスク装置のヘッド位置等の状態を動的に判断して、その論理ディスク装置への入出力命令実行時間が最短となるように入出力命令の発行をするようにしたものである。

【００２３】第１２の発明に係わる入出力処理システムは、前記論理ディスク制御手段が、複数のディスク装置の各々に複数回の入出力要求が発生する入出力サイズで論理ディスク装置に対して入出力要求がある場合は、この論理ディスク装置を構成するディスク装置の性能特性、および、ディスクが接続された入出力バスの特性を判断し、必要に応じて、同一ディスク装置への複数入出力要求をより少ない入出力要求回数とする、あるいは、該複数ディスク装置への入出力要求中に同期ポイントを設けることにより、該論理ディスク装置への入出力時間が最短となるように入出力命令を発行するようにしたものである。

【００２４】第１３の発明に係わる入出力処理システムは、前記論理ディスク制御手段が、システム管理者より与えられた性能を満たす論理ディスク装置を自動的に構成するようにしたものである。

【００２５】第１４の発明に係わる入出力処理システムは、入出力装置に対するデータ転送制御を行うデータ転送ドライバを備え、このデータ転送ドライバは入出力命令の入出力対象装置に論理ディスク装置が含まれている場合は、システムメモリ上にデータ転送用の2面バッファを確保してデータ転送を行うようにしたものである。

【００２６】第１５の発明に係わる入出力処理システムは、入出力装置の接続される入出力バスの制御をするバスアービタを設け、このバスアービタはデータ転送元及びデータ転送先の装置のＩＤを登録するソースレジスタ及びデスティニレジスタを備え、前記データ転送ドライバは、入出力命令の入出力対象装置に論理ディスク装置が含まれていない場合は、前記バスアービタを駆動させることによりシステムメモリを利用せずにデータ転送を行うようにしたものである。

【００２７】第１６の発明に係わる入出力処理システムは、前記データ転送ドライバが、入出力命令で指定された入力装置と出力装置のデータ転送速度に差がある場合は、システムメモリ上にデータ転送用のバッファを確保して前記バスアービタを駆動するようにしたものである。

【００２８】第１７の発明に係わる入出力処理システムは、前記論理ディスク装置を構成する複数のディスク装置が各々異なる入出力バスに接続されて複数の論理ディスク装置を構成する入出力処理システムにおいて、前記

各入出力バスに接続された各々のディスク装置を制御するディスク制御装置に同じ入出力バスに接続されたディスク装置間での複写手段を設け、異なる論理ディスク装置間でデータの複写を行うようにしたものである。

【００２９】

【作用】第1の発明に係わる入出力処理システムにおいては、性能データ収集手段がシステム管理者が提供する入出力装置の性能データもしくは性能計測手段が実際に入出力データを動作させることにより得る性能データを収集する。得られた性能データを基に論理ディスク構成手段が論理ディスクを構成する各ディスク装置の1ストライプ分のデータ入出力に要する応答時間が均等になるように論理ディスクを構成するための論理ディスク構成データを生成する。こうして得られた論理ディスク構成データを用いて論理ディスク制御手段が論理ディスクを制御する。

【００３０】第2の発明に係わる入出力処理システムにおいては、制限時間設定手段により入出力命令出指定させた入出力装置の動作に要する時間を設定し、設定された制限時間が経過すると入出力命令を終了させる手段により入出力動作を未完了状態であっても終了させる。

【００３１】第3の発明に係わる入出力処理システムにおいては、入出力装置の制御をする制御装置に設けられたタイマに制限時間を設定し、このタイマから制限時間が経過したことを通知されると同じく制御装置に設けられた入出力動作を終了させる手段により入出力動作をたとえ実行中であっても終了させる。

【００３２】第4の発明に係わる入出力処理システムにおいては、複数のディスク装置で構成された論理ディスク装置に対する入出力命令の発行時に制限時間設定手段により入出力動作に要する制限時間を設定し、設定された制限時間を経過すると入出力動作を終了させる手段により動作を終了させる。

【００３３】第5の発明に係わる入出力処理システムにおいては、第1の発明により構成された論理ディスク装置に対する入出力命令の発行時に制限時間設定手段により入出力動作に要する制限時間を設定し、設定された制限時間を経過すると入出力動作を終了させる手段により動作を終了させる。

【００３４】第6の発明に係わる入出力処理システムにおいては、入出力装置を実際に動作させて性能を計測する性能計測手段が、性能を測定する際に、入出力命令に対する応答時間が最短、あるいは、最長となる条件を設定して入出力命令を実行して応答時間（性能）を計測する。

【００３５】第7の発明に係わる入出力処理システムにおいては、入出力装置を実際に動作させて性能を計測する性能計測手段が、システムに初めてディスク装置が導入される時には必ず必要な初期化に合わせて、初期化の一部として性能を計測する。

【００３６】第8 の発明に係わる入出力処理システムにおいては、性能収集手段が入出力装置の性能の計測に加えて、入出力装置の接続されたバスの性能をシステム管理者より与えられるバスの性能データもしくは性能計測手段により実際に入出力装置を動作させることにより得る性能データを収集し、論理ディスク構成手段が得られた性能データを基にして論理ディスク構成データを生成する。

【００３７】第9 の発明に係わる入出力処理システムにおいては、論理ディスク構成手段が論理ディスクを構成する各ディスク装置に対する１回の入出力命令のデータ転送量が各ディスク装置の1トラックの中に収まるように論理ディスク構成データを生成する。

【００３８】第１０の発明に係わる入出力処理システムにおいては、論理ディスク構成手段が、性能収集手段で収集した性能データを参照した結果、ディスク装置の内周部と外周部で性能に差があると判断した場合には、各々のディスク装置の外内周部を交互に組み合わせた構成の論理ディスク装置を構成する。

【００３９】第１１の発明に係わる入出力処理システムにおいては、論理ディスク制御手段が、論理ディスク装置をアクセスする場合に、論理ディスクを構成する各ディスク装置のその時点のヘッド位置を認識して、入出力命令としての実行時間が最少となるようにディスク装置への起動順序を決定する。

【００４０】第１２の発明に係わる入出力処理システムにおいては、論理ディスク制御手段が、論理ディスクを構成する各ディスク装置に対する入出力命令が複数回の入出力命令に分割しなくてはならないサイズで要求されたときは、論理ディスク構成データを参照して入出力装置に対する入出力命令数が前記分割数よりも少ない回数で済むように制御する。

【００４１】第１３の発明に係わる入出力処理システムにおいては、論理ディスク構成手段が、論理ディスク構成データを自動的に生成する。

【００４２】第１４の発明に係わる入出力処理システムにおいては、データ転送ドライバが、入出力命令の対象とする入出力装置に論理ディスク装置が含まれている場合には、システムメモリ上に2 面バッファを確保して、アプリケーションプログラムではデータ転送バッファを意識せずに済むデータ転送を行う。

【００４３】第１５の発明に係わる入出力処理システムにおいては、データ転送ドライバが、入出力命令の対象とする入出力装置に論理ディスク装置が含まれていない場合には、データ転送元及びデータ転送先の入出力装置のＩＤをバスアービタの対応するレジスタに登録する。以降、バスアービタは、入力側装置と出力側装置とにバスの使用権を与えてシステムメモリを使用せずにデータ転送を行う。

【００４４】第１６の発明に係わる入出力処理システム

においては、データ転送ドライバが、入出力命令の対象とする入出力装置の論理ディスク装置が含まれていなくて、且つ入力側装置と出力側装置の転送速度に差があると判断すると、システムメモリ上にデータ転送用のバッファを確保した上で、バスアービタを駆動する。

【００４５】第１７の発明に係わる入出力処理システムにおいては、ディスク制御装置に設けられた複写手段により同一ディスク制御装置に接続されたディスク装置間での複写を行う。

【００４６】

【実施例】

実施例1．図1 は本発明を適用したハードウェアシステムの構成を示す図である。図において、１０１は本システム全体の処理をするＣＰＵ、１０２はＣＰＵ１０１が処理する命令及びデータが格納されるメモリ、１０３は時間を計測するシステムタイマで、各々メモリバス１０４に接続されている。メモリバス１０４はブリッジ１０６を介してローカルバス（主たる入出力バス）１０７に接続されている。このローカルバス１０７はバスアービタ１０５により制御される。また、ローカルバス１０７には、ＤＭＡコントローラを備えたＳＣＳＩバスアダプタ１０９を介して同じ構成の4 式のＳＣＳＩバス１１２と高速デバイス１０８が接続されている。この高速デバイス１０８は、ＤＭＡインターフェースを有する装置で、例えば、ソートプロセッサのようにディスク装置から受け取った大量のデータに対して即時的なデータ処理を行い、処理結果をディスク装置へ再び出力するものである。さらに、各ＳＣＳＩバス１１２には、ディスクコントローラ１１０を介して同じ構成の4 台のディスク装置１１１が接続されている。本実施例では、ディスク装置１１１は同一のものが使用されているが、もちろん、異種のものが接続されても構わない。

【００４７】図2 は本発明のソフトウェア構成を示す図である。図において、２０１はディスク装置１１１、または、論理ディスク装置から、高速デバイス１０８へ、あるいは、逆に高速デバイス１０８からディスク装置１１１または論理ディスク装置へデータ転送を行う制御プログラムである。この制御プログラム２０１は1 ユーザアプリケーションプログラムとして動作する。２０２はオペレーティングシステムであるが、本実施例においては、オペレーティングシステム２０２はフレームワークとして主に利用され、オペレーティングシステム自体の機能はほとんど使用されない。制御プログラム２０１も実質的にはデータ転送ドライバ２０３に制御を依頼する。つまり、「デバイスＡのオフセット１MB 目から３０MB のデータをデバイスＢに書き込め」という形で指令を出すのみである。データ転送ドライバ２０３は、その名の通り、疑似ドライバとしてカーネルに組み込まれる。２０４はソフトウェア制御によって、複数のディスク装置を使用して論理ディスク装置を構成して、オーバ

11

ーヘッドを排して、より小さい応答時間、つまり、より高速な大量データ転送、あるいは、ある制限時間内での確実な応答を得るためのソフトウェアである（このソフトウエアが論理ディスク制御手段の一部で、以後、このソフトウェアをボリュームマネージャと呼ぶ）。２０５はディスク装置１１１のコントロールを行うディスクドライバ（論理ディスク制御手段の一部）、２０６は高速デバイス１０８のコントロールを行うデバイスドライバ（論理ディスク制御手段の一部）である。なお、バスアービタ１０５のデータ転送機能を利用する場合のコントロールは、データ転送ドライバ２０３が直接行う。なお、論理ディスク制御手段は、論理ディスク装置を制御するだけでなく、論理ディスク装置以外の入出力装置も制御する。

【００４８】ここで、本発明の中核となるボリュームマネージャ２０４の機能について説明する。ボリュームマネージャ２０４は、複数のディスク装置により論理ディスク装置を構成することにより、実際のディスク装置よりも大きな連続データ領域を持つ論理ディスク装置を提供したり、論理ディスク装置への書き込みを複数のディスク装置に反映することによるディスクミラーリングや、ストライプ構成された論理ディスク装置を構成して、入出力性能の高速化等を行う。つまり、ボリュームマネージャ２０４は自身の管理下におかれたディスク装置からシステム管理者の指示によって論理ディスク装置を構成し、ユーザは論理ディスク装置に対して、アクセスを行い、ボリュームマネージャ２０４の管理するディスク装置に直接アクセスすることはない。もちろんシステム管理者は全てのディスク装置をボリュームマネージャ２０４の管理下に置く必要はない。

【００４９】ボリュームマネージャ２０４について、もう少し説明をしておく。ボリュームマネージャ２０４は、ディスク装置をシステムに導入するディスク導入手段２０７、このディスク導入手段２０７がディスク導入時にディスク装置から読み取ったディスクラベルとシステムのディスク属性ファイルから生成されるディスク管理データ２１０、このディスク導入手段が起動するディスク装置の性能を測定する性能計測手段、この性能計測手段で測定した性能データ２０９及び、この性能データ２０９とディスク管理データ２１０とから論理ディスク構成データを生成する論理ディスク構成手段２１３、論理ディスク構成手段２１３で生成された論理ディスク構成データを用いて論理ディスク装置へのアクセス時のインタフェースを司どる論理ディスク入出力インタフェースから構成される。

【００５０】ここで、ディスクストライピングという技術につき整理しておく。理解を容易にするために、２台の全く均等なディスク装置が、十分に高速な入出力バスに接続されているとすると、あるサイズの連続した大量データは、例えば１２８セクタ分のサイズ（以後、このデータ量単位をチャンクと呼ぶ。）に区切られ、チャンクは1番目から交互に、各々のディスク装置に順番に格納される。このようにストライプ構成された論理ディスク装置の先頭に2つのチャンクを含む入出力サイズで入出力要求を発行すると、入出力要求は2つのディスク装置に振り分けられ処理されるので、見かけの入出力性能は2倍になる。

12

【００５１】さて、例えば、システム管理者が通常のディスク装置の2倍の性能のディスク装置が必要となり、通常のディスク装置2台とボリュームマネージャ２０４を利用してこれを実現する場合で説明する。システム管理者は、ボリュームマネージャ２０４をシステムに導入する。ボリュームマネージャ２０４は、ディスク導入手段２０７に対して、新規2台のディスク装置の導入指示を与え、ボリュームマネージャ２０４の管理下に登録する。登録要求を受けるとボリュームマネージャ２０４はディスク装置１１１の初期化を行い、論理ディスク装置とするための管理メタデータ（ディスク管理データ２１０）を書き込む。この状態では、まだ、論理ディスク装置としての構成は、なされていない。システム管理者はさらに、ボリュームマネージャ２０４に2台でストライプ構成の論理ディスク装置を構成するように、論理ディスク構成手段２１３に依頼する。論理ディスク構成手段２１３は論理ディスク構成データ２１０を生成し、これによって初めてユーザは論理ディスク装置を利用できるようになる。ユーザが論理ディスク装置にアクセスする場合は、論理ディスク入出力インターフェース２１２にオペレーティングシステム２０２を経由してアクセスする。入出力要求を受けるとボリュームマネージャ２０４は論理ディスク装置の論理ディスク構成データを参照して、論理ディスク装置を構成する実際のディスク装置１１１に対し、入出力命令を出す。

【００５２】以下、順次本入出力処理システムの動作につき説明していく。最初に、システムへのディスク装置を導入する時に行われる、ディスク装置単体の性能の計測、バス能力の計測等について説明する。先ず、概要について説明する。これは、ディスク装置等の性能を性能計測手段２０８（性能データ収集手段の一部）によって、ディスク装置の入出力サイズによる応答性能、ディスクアドレスによる応答性能、ディスクキャッシュ効果等の推定を行い、ストライプを構成する各ディスク装置の性能特性を知るものである。具体的には、上記パラメータを用いてディスク装置に対して、入出力動作を実行させてその応答時間（性能データ２０９の一部）を測定することによって行い、その結果を記録する。そして、その測定した各ディスク装置の性能データ２０９を基に、ディスクストライプを構成するディスク装置の1ストライプ分のデータ入出力に要する応答時間が各々均等になるように各ディスク装置に配置すべきデータ量、つまり、ストライプ幅を決定し、それに沿った論理ディス

13

ク構成データ２１１を論理ディスク構成手段２１３により、即ちストライピング構成の論理ディスク装置を構成する。こうして構成された論理ディスク装置に対しての入出力要求は、ストライプを構成する各々のディスク装置に振り分けられ、各々の入出力装置は均等なある時間で応答するので、大まかには、単純にディスク装置に対して入出力する場合に比較して、見かけの応答時間はストライプを構成するディスク装置台数分の1となる。

【００５３】ボリュームマネージャ２０４は、その後にシステム管理者によって要求される論理ディスク装置構成要求時にこの計測された性能データ２０９を基に、要求された論理ディスク装置性能を確保できるかどうかを判断する。ディスク装置の導入はディスク装置接続の後システムを立ち上げ、ディスク接続確認後、フォーマット（初期化）を行うコマンド投入によって行われるが、書き込みデータとしてフォーマット用データを用いることにより、ディスク性能測定とディスク装置の初期化を兼ねる。これは、単なるフォーマットではなくて、ボリュームマネージャ２０４の管理下に置くための、メタデータ（ディスク管理データ２１０）が書き込まれる。図３に示すのはディスク装置１１１に対して行った性能測定の結果である。これは前記メタデータとして該ディスク装置上に格納されたものをそのままきり出したスナップショットである。

【００５４】図3において、５０１はシステムの中で該ディスク装置１１１を特定する名前、５０２は、このディスク装置１１１の型式を特定するインデックス情報である。ボリュームマネージャ２０４はディスク装置１１１の諸元についてのカタログデータを持ったデータベースをシステム管理者から与えられて、あるいは、ディスク装置１１１のラベル情報を読み出すことにより、この情報を得る。これにより、ディスク回転数、各シリンダにおける、1トラックあたりのセクタ数などを知ることができる。５０３はディスク装置１１１の容量、５０４はシーク、回転待ちなしの場合のアクセスタイム、５０５は最も遠い位置からシークし、１回転待った場合のアクセスタイムで、マイクロ秒で記されている。この計測は、基準位置へコマンド発行した後、次セクタ、あるいは、最も遠いセクタへのアクセス命令を発行し、基準位置へのアクセス完了から、計測位置へのアクセス完了までの経過時間を測定することによって行う。５０６から５０７までの１０個のデータは0シリンダから、１００MB毎の位置で転送速度を計測したものがＫＢ／Ｓで記されている。これらのディスク性能データ２０９は後に説明する論理ディスク構成の際に参照／利用される。なお、本実施例で使用したディスク装置は読み出し性能と書き込み性能に差がないものとする。

【００５５】一般に、システムにディスク装置を新規に導入する時には、不良セクタの発見や、ラベル情報の書き込みのために、フォーマットと呼ばれる処理を行う

14

が、この実施例ではこのフォーマット時に、上記のように性能計測手段２０８によりディスク装置の性能データ２０９を収集する作業を兼ねて行うことができるにしているので、性能計測に要するコストを低減でき、システム管理者の負担を低減できる。

【００５６】次に、入出力バスのスループット性能の測定について説明する。この測定は、ディスク装置の入出力性能や同期動作を妨げる要因となり得る入出力バスのデータ転送性能（性能データ２０９の一部）や同一バス上での競合を考慮して、論理ディスクの構成を行うためのものである。システム管理者により与えられた、入出力バスのデータ転送能力及び構成データから、もしくは、複数の入出力バスに各々接続された全てのディスク装置を含めた入出力装置を組み合わせて駆動し、負荷をかけ、各々の入出力装置の入出力動作の応答時間を測定することにより、各々の入出力バス、さらにそれらが接続された上位の入出力バスのデータ転送能力に余裕があるか、あるいは、転送能力の限界値はどれほどかを知り、ストライプ構成の論理ディスク装置への入出力時に、並列動作するディスク装置のデータ転送により共有する入出力バスのデータ転送能力の限界となることを回避するように、論理ディスク装置をストライプ構成するディスク装置を選択して構成することにより、論理ディスク装置への入出力時に入出力バスネックとなることによる性能劣化を回避するためのものである。

【００５７】従って、この測定は、単なるスループット性能だけでなく、入出力バス競合の発生条件を測定対象バスに接続された入出力装置に対して、起動する順番によって規定して、実際にバス競合を発生させ、各々の入出力装置の入出力レイテンシを測定するものである。これにより、バスアービトレーションポリシー等を知り、ストライプを構成するディスク装置の入出力起動時に不要なバス競合を避けるための、起動遅延を挿入するといった適用が可能であるが、本実施例では、ディスク装置が接続されるＳＣＳＩバス１１２、ローカルバス１０７のスループット性能を計測して、ストライプ構成される論理ディスク装置に対する入出力時に、各入出力バスのスループット性能の限界を超えないようにストライプを構成する。ＳＣＳＩバス１１２については、接続された入出力装置、本実施例では、たまたま、同一のディスク装置4台が全てのＳＣＳＩバス１１２に同様に接続されているが、該バスに接続された、入出力装置を全て起動して一定時間に得た入出力データの総量を計算することにより、スループット性能の限界値を知る。これを、ディスク装置一つを動作させた場合のスループット性能と比較して、何台のディスク装置まで、バス転送速度の制約を受けずに動作可能かを知る。本実施例の場合ディスク装置の最大転送速度が2 MB／s で、ＳＣＳＩバス１１２の最大転送速度は8 MB／s であるのでＳＣＳＩバス１１２に接続された4台のディスク装置を同時に動作

15

させても、単純にはバス転送速度の制約を受けないことがわかる。

【００５８】同様に、ローカルバス１０７の転送速度を全てのディスク装置、及び、高速デバイス１０８を動作させることにより計測したところ、４０MB／s で、全ての入出力デバイスの転送速度の合計に相当し、結局、本実施例の入出力バスと入出力装置の構成においては、全ての入出力装置に対してどのように入出力動作を実行しても入出力バスの転送速度による制約を受けないことがわかる。

【００５９】ところが、本実施例とは異なり、例えば、ＳＣＳＩ バスの転送速度が本実施例のＳＣＳＩ バスの転送速度より遅く、そのＳＣＳＩ バスの転送速度では2 台分のディスク装置しか、ディスク装置の最大転送速度では動作できないのであれば、ストライプ構成で同一の論理ディスク装置を構成するディスク装置を3 台以上、同一ＳＣＳＩ バス上に接続されたディスク装置では構成しないようにすることになる。

【００６０】次に、論理ディスク装置の構成方法について説明する。図4 は、ストライプ構成の論理ディスク装置を構成する論理ディスク構成手段２１３の動作を説明するフローチャートである。以下、この図を参照しながら説明する。ステップ６０１において、ユーザ、つまり、入出力装置を管理するシステム管理者は、必要とするストライプ構成の論理ディスク装置の性能、例えば、８MB／Sと、アプリケーションプログラムが論理ディスク装置のアクセスの際に発行する入出力サイズ、例えば、５１２KB、及び、必要とする論理ディスク装置のサイズ、例えば、１GBを入力する。これにより、本入出力処理システムは、指定された仕様を満たす論理ディスク装置の構成を試みる。ここで、入出力サイズを指定するのは、ストライプ数、及び、ストライプ幅が入出力サイズによって制限されるためである。つまり、入出力サイズは、上述のチャンクサイズの倍数でないと、ディスクストライピングによる性能向上が得られないためである。

【００６１】次に、ステップ６０２において、性能計測手段２０８によって得たディスク装置の性能データ２０９から、指定された論理ディスク装置の性能を達成するのに何台のディスク装置の並列動作が必要か、つまり、ストライプのWAY数を決定する。WAY数が決定したら、ステップ６０３で本入出力処理システムの管理するディスク装置上の空き領域を探し、必要なWAY数分の異なるディスク装置上に、ストライプを構成可能かどうか調べる。ただし、指定された性能基準を満たすのに、ディスクストライピングの必要のない場合もあるので、この時は、単にディスク装置上の空き領域を探すことになる。その後ステップ６０４で、与えられた入出力サイズによって、この出入力サイズをチャンクサイズとするストライプ構成の論理ディスク装置を構成した場合に、

16

性能上の効果があるか、あるいは、要求された性能水準はディスクストライピングを必要としないのかを判断し、ストライプ幅の制約により効果がない、あるいは、ディスクストライピングの必要のない場合は、ディスク領域を確保の上期待できる論理ディスク装置性能を、ステップ６１０でユーザに報告して終了する。

【００６２】２WAY以上のディスクストライピングが有効な場合はさらに、ステップ６０５で同種のディスク装置でストライプ構成が可能か否かを判断し、構成することが可能な場合には、ステップ６０６でこのディスク装置が内外周部で性能差があるか否かを判断し、内外周部で性能差がある場合には、ステップ６０７でさらにストライプを構成するディスク装置が同一のオフセット位置から使用可能か否かを判断する。そして、同一のオフセット位置から使用可能であれば、ステップ６０８でディスク装置を交互に逆方向にストライプ構成するようにする。また、ステップ６０５またはステップ６０６またはステップ６０７で否と判断したときは、ステップ６０８でストライプ幅で個々のディスク装置の応答速度を平均化するようストライプ構成をする。

【００６３】ステップ６０９における動作を以下に補足して説明する。ディスクストライピングは次のように構成される。まず、チャンクサイズを決定するが、指定された入出力サイズが、論理ディスク装置の性能を達成するのに必要なストライプ幅の総計より大きくない限り、入出力サイズがそのままチャンクサイズとする。一方、入出力サイズが大きすぎる場合は、必要に応じて、入出力サイズの自然数分の1 をチャンクサイズとする。いま、入出力サイズとして５１２KBが指定されているとする。この場合、例えば、ディスク装置の性能データ２０９から、平均３２KBストライプ幅の4 WAYで性能達成が可能と考えるとチャンクサイズは、５１２KBの１／４の１２８KBとする。すでに取得しているディスク装置の性能データ２０９から、ストライプを構成するディスク装置の当該位置のデータ転送速度を得て、その、逆数の比で、当該チャンクを分割して、各々のストライプ幅とする。これにより、該論理ディスク装置に対する入出力動作の実行時には、各々のストライプを構成するディスク装置の応答時間は平均化される。

【００６４】次に、ステップ６０８における動作を補足して説明する。ステップ６０８におけるディスクストライピングは次のように構成される。この場合も最初にチャンクサイズ、及び、各ディスク装置のストライプ幅が決定されるが、偶数個用意されるストライプ構成されたディスク装置の同一の使用領域の先頭からと、最後尾から交互にストライプが構成され、そのストライプ幅は1 トラック分とする。ディスク装置の性能情報の５０２により示される当該位置のトラックの含むセクタ数より、該先頭位置と最後尾位置の1 トラックのサイズを知り、（先頭トラックのサイズ＋最後尾トラックのサイズ）×

17

ストライプWAY数／2 がチャンクサイズとなる。従って、ストライプ幅は各々の入出力時点でトラックあたりのサイズになる。

【００６５】以上のように、論理ディスク装置を構成するための論理ディスク構成データ２１１を生成登録したならば、推奨する入出力サイズ、つまり、内部で決定されたチャンクサイズと期待できる、この論理ディスク装置の性能をステップ６１０で報告して終了する。あらかじめ与えられた性能水準と入出力サイズにより、ディスクストライピングの必要がない場合、あるいは、効果がない場合は、推奨する入出力サイズとして、単体ディスク装置に対して、与えられた性能水準を満たすのに必要な入出力サイズを報告し、効果がない場合は、期待できる性能として単体ディスクの性能を報告する。

【００６６】図5 、図6 は以上のようにして構成された論理ディスク装置、及び、論理ディスク装置を構成する各ストライプ構成されたディスク装置の管理情報の一部である。これら管理情報は本入出力処理システムの管理する全てのディスク装置１１１上にメタデータとしてそのコピーが格納され、必要に応じて、本入出力処理システムの管理するメモリ１０２上に読み出される。便宜のため図には一般的な表形式で書かれているが、実際のディスク装置、メモリ上には、プログラムで管理されたテンプレートに従い格納される。

【００６７】以下、図5 、図6 に示す、ステップ６０８の実行により構成された論理ディスク装置及びストライプ構成されたディスク装置の管理情報について説明する。図5 において、７０１は該論理ディスク装置のデバイスノード名である。アプリケーションプログラムはこのノードをディスク装置としてアクセスする。７０２は該論理ディスクのサイズ、つまり、容量を示している。７０３は、該論理ディス装置の構成情報として4 WAYのストライプ構成であることを示し、７０４は、ストライプを構成するディスク装置とそのディスク装置の領域を示していて、各々、ｃ０ｔ０ｄ０－０、ｃ１ｔ０ｄ０－０、ｃ２ｔ０ｄ０－０、ｃ３ｔ０ｄ０－０という名称で特定されている。理解を容易にするために、ｃ０ｔ０ｄ０がディスク装置を特定し、－０が領域番号を特定するようにしているが、これは論理的なものであるので、ハードウェア上のアドレスを特定するものではない。また、図6 には、７０４に示すディスク装置のの領域情報、つまり、リージョン情報へのインデックスｃ０ｔ０ｄ０－０によって特定されるリージョン情報が示されている。８０１は該リージョン名、８０２は該リージョンが配置される物理ディスク装置名、８０３はリージョンの開始オフセット、８０４は該リージョンのサイズ、８０５は該リージョンにおけるデータの配置方向、８０５には該リージョンにおけるストライプ幅を示している。

【００６８】次に、図5 、図6 に示される管理情報（論理ディスク構成データ２１１）を持つ論理ディスク装置

18

に対して入出力動作を要求された時の動作につき説明する。アプリケーションプログラム（制御プログラム２０１もアプリケーションプログラムの1 つである）から本論理ディスク装置に対してアクセスする時は、入出力サイズとしてチャンクサイズ、及び、論理ディスク装置に対するオフセット値が与えられる。言い替えると、アプリケーションプログラムはあらかじめ、ボリュームマネージャ２０４に対してチャンクサイズを問い合わせることができる。ボリュームマネージャ２０４は問い合わせ要求を受けると指定された論理ディスク装置のチャンクサイズ等の属性を報告する。本入出力処理システムは、入出力要求を受けとると、論理ディスク装置の管理データ２１１を参照して、この論理ディスク装置は4 WAYのストライプ構成で、７０４で示されるリージョンから構成されていることを知る。また、リージョン情報からこの論理ディスク装置がストライプ構成されていることを知ると、入力された論理ディスク装置に対するオフセット値から、このオフセットはストライプ何チャンク分に当たるかを計算し、ストライプを構成する各々のディスク装置に対し、配置方向がリバースである、つまり、逆方向のリージョンについては、該リージョンの最後尾からオフセットチャンク数分のトラックサイズの累計をオフセットとし、該トラックのトラックサイズを入出力サイズとして、該ディスク装置に入出力命令を発行する。順方向のリージョンに対してはリージョンの先頭から、同様に、オフセットを計算し、該トラックのトラックサイズを入出力サイズとして入出力命令を発行する。論理ディスク装置を構成する全てのリージョンについて、ディスクドライバ２０５の保持する該ディスク装置のヘッド位置から、最もヘッド位置の遠いものから入出力を起動し、その全ての入出力の完了を待つことによって、この論理ディスク装置への入出力動作は終了する。ただし、リージョン内の各々の位置のトラックサイズはすでに説明したように、本入出力処理システムがディスク装置の種類毎に備えるデータベースによって得る。

【００６９】以上のように、この実施例によれば、システムに接続されているディスク装置の性能特性を性能収集手段がシステム管理者の与えるデータもしくは性能計測手段により収集し、収集したデータを基に、論理ディスク構成手段により論理ディスクを構成し、論理ディスク制御手段により論理ディスクを動作させるので、論理ディスクを構成する各ディスクの1 ストライプ分のデータ入出力に要する応答時間が均等となり、論理ディスク装置の性能向上を図ることができる。

【００７０】また、システムにディスク装置を新規に導入する時には、不良セクタの発見や、ラベル情報の書き込みのために、必ず実行されるフォーマット処理を兼ねて、性能計測手段２０８によりディスク装置の性能データ２０９を収集する作業を行うことができるにしているので、性能データ２０９の収集に要するコストを低減す

19

ることができると共にシステム管理者の負担を低減できる。

【００７１】また、実際にシステムに装着されているディスク装置各々の、最大アクセス時間、最小アクセス時間、及び、転送速度を、実際に個々の装置に対して入出力動作を実行し、その応答時間を測定して、測定したデータを一覧表にしたテーブルを用いて論理ディスクを構成するので最適な論理ディスク装置を構成できる。

【００７２】さらに、ディスク装置の入出力性能や同期動作を妨げる要因となり得る入出力バスのデータ転送性能や同一バス上での競合を考慮して、論理ディスク装置を構成するために、システム管理者により与えられた入出力バスのデータ転送能力及び構成データを用いて、もしくは、性能計測手段により複数の入出力バスに接続された全てのディスク装置を含めた入出力装置を組み合わせて動作させることにより、負荷をかけ、各々の入出力装置の入出力動作における応答時間を測定することにより、各々の入出力バス、さらにそれらが接続された上位の入出力バスのデータ転送能力に余裕があるか、あるいは、転送能力の限界値はどれほどかを知り、ストライプ構成の論理ディスク装置への入出力時に、並列動作するディスク装置のデータ転送により共有する入出力バスのデータ転送能力の限界となることを回避するように、論理ディスク装置をストライプ構成するディスク装置を選択して構成するようにしたので、構成された論理ディスク装置に対する入出力動作時に入出力バスネックとなることによる性能劣化を回避することができる。

【００７３】また、論理ディスク装置のストライプを構成する１チャンク分のデータをディスク装置上の1トラックに収まるように配置することにより、ヘッドシークのコスト（位置決めに要する時間）をシーケンシャルアクセスにおいて、その方向によらず平均化する手段を設けるようにしたので、外周から内周方向へシークしても、内周から外周方向にシークしても、読みだし時のディスクキャッシュ効果を除き差がなくなるので、より柔軟にディスクストライプを構成できる。

【００７４】さらに、内外周部で性能差のある複数のディスク装置でストライプ構成する場合には、内外周部が交互に組み合わされるようにストライプを構成することにより、ストライプ構成された論理ディスク装置の読みだし／書き込み位置に関わらず、均一な入出力応答時間を得ることができる。

【００７５】さらに、ストライプ構成された論理ディスク装置に対して入出力要求があった場合に、そのアクセス要求位置へのディスクヘッドの現位置からの距離と各ストライプを構成するディスク装置の性能的に有利な入出力サイズを考慮して、その論理ディスク装置としての入出力動作が最も短時間に終了するように、全ての必要とされるディスク装置への入出力命令の発行順序を制御するようにしたので、論理ディスク装置としての入出力

20

動作の効率が向上する。

【００７６】さらに、論理ディスク装置を構成するWAY数、ストライプ幅等の要素を自動的に生成するようにしたので、システム管理者の負荷を軽減し、容易に最適な論理ディスクを構成できる。

【００７７】実施例2．本実施例は、実施例1で説明したようにして構成された論理ディスク装置に対して、複数のチャンクを含む入出力サイズで入出力要求があった場合、同一のディスク装置に対する複数の入出力動作を取りまとめて行うものである。この実施例は、ストライプ構成された論理ディスク装置に対して複数のチャンクを含む入出力サイズで入出力要求があった時、同一ディスク装置に対して発行される複数回の入出力命令ををを１回にまとめた方が、その論理ディスク装置への入出力動作がより速く終了するならば、全体性能として、論理ディスク装置の入出力性能を向上させることができることを前提にするものである。しかしながら、論理ディスク装置を構成するディスク装置、及び、その接続された入出力バスの優先度決定方式、あるいは、ディスク装置をドライブするデバイスドライバの特性によって、必ずしも、ディスク装置への入出力命令をとりまとめて発行することが、全体性能の向上につながらないことがあり、このような場合においては、必要に応じて積極的にディスク装置間の動作を同期する、同期ポイントを挿入することにより、論理ディスク装置に対する入出力時間の向上を図る。

【００７８】本実施例においては、複数の入出力命令を纏めて１つの入出力命令とするので、所謂データチェーニング機能を必要とするが、ディスクドライバ２０５、及び、ＳＣＳＩバスアダプタ１０９に、メモリとのデータ転送時のスキャッタリング／ギャザリング機能を持たせ、ＤＭＡリストにより、このデータチェーニングを実現している。

【００７９】実施例1で説明した論理ディスク装置では、シーク方向が逆になるデータ配置方向のリージョンを含んでいるので、必ずしも入出力命令ををまとめることが性能的に有利とならないため、この入出力命令の取りまとめは行わない。もし、実施例1の論理ディスク装置が図4で説明したステップ６０９の実行により構成された論理ディスク装置であれば、本実施例において、入出力バスは転送能力に十分余裕を持つので、有利であると判断され、同一ディスク装置上の複数の入出力命令の取りまとめが行われる。

【００８０】実施例3．　本実施例は、本発明の入出力処理システムにおける高速処理を実現するためのさらに他の実施例である。図7はその実現手段をディスクコントローラ１１０上に設けた構成を示すブロック図である。図において、３０１はＳＣＳＩ制御部、３０４はディスク制御部、３０５はタイマ部である。本実施例においては、ディスクコントローラ１１０上にタイマ部３０

５を設けるようにしたが、メモリバス１０４に接続されたシステムタイマ１０３を用いることにより実現することも可能である。

【００８１】まず、本実施例におけるディスクドライバ２０５とディスクコントローラ１１０の動作について説明する。ディスクコントローラ１１０はＳＣＳＩベンダユニークコマンド（制限時間設定手段の一部）によりディスクドライバ２０５から、ＳＣＳＩバスアダプタ１０９、ＳＣＳＩインターフェース３０１、ＳＣＳＩ制御部３０２を経由して、タイマ部３０５（制限時間設定手段の一部）にセットするタイムアウト値を受け取ることができる。該コマンドによってタイマ中のカウンタレジスタ（図示せず）にタイムアウト値がセットされた後、リード、または、ライトコマンドをイニシエータ（コマンドの発行元）から受け取ることを契機にカウントダウンを始める。カウンタレジスタが0になるとタイマ部１０５はカウントダウンを停止し、ＳＣＳＩ制御部３０２に対してカウント値が0になったことを報告する。この報告を受けるとＳＣＳＩ制御部３０２はコントローラ内部バスのアボートラインをアサートする。また、ディスク制御部３０４はいかなるフェーズであっても、ＳＣＳＩ制御部３０２から参照可能な内部レジスタ（図示せず）に、中断時フェーズ、転送済みデータ量などを記録して動作を中止する。ＳＣＳＩ制御部３０２は、ＳＣＳＩバス１１２上でディスコネクトされていればイニシエータとリコネクトの上ステータスフェーズに遷移して、転送済みデータ量（ステータス情報）をメッセージとして送出する。なお、全てのデータ転送完了時になおタイマ動作中であれば、ＳＣＳＩ制御部３０２がカウンタレジスタを0にリセットする。

【００８２】次に、本実施例の詳細な動作を説明する。ディスクドライバ２０５はタイムアウト値を設定時に次のように動作する。ディスクドライバ２０５の使用者プログラム、即ち本実施例におけるデータ転送ドライバ２０３、あるいは、ボリュームマネージャ２０４は、ディスクドライバ２０５にデータ転送時のタイムアウト動作を要求するときには、ディスクドライバ２０５のデータ転送サブルーチンコール時に、ディスクドライバ２０５が自由に利用できる、システムに共通なブロックデバイスコントロールテーブル（ディスク装置のようにブロック単位で転送を行うデバイスを制御するためのテーブル）のフィールドを使用してタイムアウト値をセットする。タイムアウト値がセットされるとディスクドライバ２０５はＳＣＳＩバスアダプタ１０９に対し、リード／ライトコマンドの発行を要求する前に、該タイムアウト値を渡し、上述のベンダユニークコマンドの発行を要求する。続いて、リード／ライトコマンドの発行を要求するとディスクドライバ２０５は、ＳＣＳＩバスアダプタ１０９から処理の終了を割り込みによって通知される迄スリープ状態となる。処理終了の通知を受けると、ディスクドライバ２０５は、上述のコントロールテーブル中のタイムアウト値を0にリセットし、転送済みのデータ転送量をコントロールテーブル中に用意されている、データ転送量（ステータス情報）として報告する。

【００８３】次に、本実施例をストライプ構成された論理ディスク装置に対して適用した具体的な例について説明する。ところで、本実施例における入出力処理システムは制御プログラム２０１によりディスク装置をストライプ構成として高速化し、高速デバイス１０８へのデータ供給、また、高速デバイス１０８からの出力を論理ディスクディスクへの書き込みを行うが、高速デバイス１０８の処理能力を有効に活用するために、特にデータ転送開始時に当たっては、実際に送られるデータ量よりも、より、処理開始までのレイテンシを小さくすることに主眼をおくものである。また、もともと大量データを扱うので、多少入出力オペレーションの回数が増えても全体性能に影響しない。

【００８４】制御プログラム２０１がデータ転送ドライバ２０３を通して、必要と判断すると、ストライプ構成の論理ディスク装置に対して、タイムアウト値と入出力サイズを与える。ボリュームマネージャ２０４はタイムアウト値を上述のブロックデバイスコントロールテーブルを通して受け取ると、ストライプを構成するディスク装置の入出力起動時に各々のディスク装置について用意されるコントロールテーブルのタイムアウトフィールドにデータ転送ドライバ２０３から受け取ったタイムアウト値を設定する。図8は4台のディスク装置でストライプ構成された論理ディスク装置に対し、タイムアウト値を設定して読み出しを実行したときに、タイムアウト発生後、ボリュームマネージャ２０４が4台のディスク装置に対して同期を取り、各々ディスク装置からデータ転送量を受け取った状態を模式的に表したものである。図８において、９０１は第1のディスク装置、９０２，９０３，９０４はそれぞれ、第2、第3、第4のディスク装置を表し、上下方向は各ディスクのデータ転送の進捗を表す。横方向の点線は該論理ディスクの1チャンク分のデータ量を表す。従って、この例においては、ストライプ構成された論理ディスク装置に対して4チャンク分が入出力サイズとして渡されたことになる。本発明による入出力処理システムにおいては、実施例2で、すでに説明したように、可能で、かつ、有効ならば複数の同一ディスク装置への入出力命令は取りまとめられる。この図8において、各ディスク装置には4チャンク分の入出力命令が取りまとめられて渡され、この、入出力命令に対してタイムアウト値が渡される。既に、説明したように、ディスクドライバ２０５、及び、ＳＣＳＩバスアダプタ１０９には、データチェーニング機能を有しているので転送先または転送元のメモリアドレスが連続していなくても問題はない（図8において平面をメモリ空間と見立てると横方向が連続である）。

23

【００８５】タイムアウト発生後、ボリュームマネージャ２０４が4台のディスク装置９０１～９０３の同期をとった後の図8に示される各ディスク装置の転送状況は、９０４に示す第4のディスク装置を除いて未完である。しかし、いずれのディスク装置も９０５で示す第3チャンクまでについてはデータ転送を完了しているのでボリュームマネージャ２０４は3チャンク分のデータ転送が完了したことを、データ転送ドライバ２０３に報告する。本実施例においては、論理ディスク装置に対して、余りに短いタイムアウト設定をするとデータ転送量が0となる可能性があることがわかる。従って利用者としては、入出力サイズとして、入出力要求を出す論理ディスク装置の複数チャンク分を与え、それに見合ったタイムアウト値を与えることが有効である。本実施例では、制限時間設定による入出力動作の終了を実施しているが、図8に示すようにほぼ全てのディスク装置のデータ転送は同期している。

【００８６】複数チャンク分の入出力サイズでリード／ライトを行った時にストライプを構成する同一のディスク装置に対する入出力の取りまとめが、その、有効性から行われない場合には、１チャンク分ずつ、ボリュームマネージャ２０４はディスク装置間の同期をとる。全てのディスク装置への入出力命令発行時には上位指定のタイムアウト値を設定するが、第2チャンク以降の入出力命令発行時には、第1チャンクの入出力命令発行から同期取得時点までの経過時間をシステムタイマにより計測し、これを該タイムアウト値から減じて、第2チャンクの入出力命令発行時にタイムアウト値として設定する。以降、該タイムアウト値が0になるまでこれを繰り返す。なお、本実施例においては、ＳＣＳＩのベンダユニークコマンドを用いて時間設定をする例を示したが、必ずしも、ベンダユニークコマンドを使う必要はなく、またＳＣＳＩインタフェースを用いる必要もない。

【００８７】以上のように、入出力動作の制限時間を設定する制限時間設定手段と設定した制限時間が経過すると入出力動作を終了させる手段を設け、指定されたデータ量の入出力動作の完了前に、指定された制限時間に達した時は、入出力動作を終了させ、その時点での転送済みデータ量を報告するようにしたので、アプリケーションプログラムの入出力同期待ちによる処理遅延を回避することができる。

【００８８】また、入出力制御装置にタイマ機能、入出力動作を終了させる手段、及び、入出力中断時のステータス報告手段（転送済みデータ量の報告）を設けるようにしたので、機能の実現が容易にできる。

【００８９】また、複数のディスク装置で構成される論理ディスク装置において、この複数のディスク装置を並列動作させる場合に、ディスク装置間の入出力完了の同期を入出力動作の制限時間を設定することによりとるようにすることにより、ユーザの希望した時間までに、こ

24

の論理ディスク装置に対する入出力処理を完了することができる。つまり、論理ディスク装置を構成するディスク装置各々に対して、上位から与えられた同じ制限時間を設定して入出力命令を発行し、その完了を待ち、全てのディスク装置から応答があった時点で、各々の、入出力動作完了ステータスをまとめて上位に報告することにより、各々のディスク装置の入出力の完了状態はまちまちである可能性があるが、ユーザの希望した時間内に論理ディスク装置に対する入出力動作を終了させることができる。

【００９０】さらに、ストライプ構成された論理ディスク装置に、複数のディスク装置間の同期方法として、この実施例で説明した同期方法を適用することにより、設定する制限時間により論理ディスク装置に対する入出力動作の応答時間は保証され、かつ、入出力動作の完了ステータスのばらつきの少ない、つまり、応答時間に見合ったデータ量設定であれば、全てのディスク装置における入出力動作が完了する可能性が高くなる。

【００９１】実施例4．　本実施例は、本発明の入出力処理システムにおける高速処理を実現するためのさらに他の実施例である。本実施例では、上記実施例1～実施例3で説明した入出力システムにおいて、入出力装置間のデータ転送を高速に行うための機構について説明する。

【００９２】従来、入出力装置間でデータ転送する時には、データ転送制御を行うユーザープログラムのデータ空間に転送元からデータを読み込み、読み込んだデータを転送先の装置に書き込むといった操作が行われていた。ユーザープログラムの有するメモリ空間は、システムのメモリ管理下におかれ、データ転送の際システムのメモリ管理にとって負担となり、場合によってユーザプログラムのデータ空間にはデバイスからＤＭＡ動作できないこともあり、これがデータ転送速度を結果的に低下させていた。これをシステムの物理メモリ上に、システムの関知しないデータ転送用のバッファメモリ領域を確保し、カーネル空間で実行される制御プロセスからコントロールして、データ転送することにより、メモリ管理へ負担を掛けることによる性能低下を防ぐことができる。

【００９３】図9は、この高速転送を、論理ディスクを含まないデバイス間での実現するためのバスアービタ１０５の例で、図において４０１、４０２は各々データ転送の転送元、転送先となる高速デバイス１０８のローカルバス１０７上のＩＤが登録される、システム上のメモリ空間にマップされたバスアービタのソースレジスタ、デスティニレジスタ、４０３はローカルバスを制御する論理回路である。図１０は、論理ディスクを含まない入出力装置間の高速転送を行うデータ転送ドライバ２０３の処理の流れを示すフローチャート図である。図１１は、論理ディスクを含む入出力装置間の高速転送を行う

25

バスアービタ１０５の処理の流れを示すフローチャート図である。

【００９４】制御プログラム２０１により、データ転送ドライバ２０３に対し、転送元デバイス、その種別、及び、データオフセット、転送先デバイス、その種別、及び、データオフセット、そして、転送データ量を指定されるとデータ転送ドライバ２０３は図１０のフローチャート図に示す論理に従って動作する。以後、図を参照しながら説明する。ステップ１００１でデータ転送ドライバ２０３は、制御プログラム２０１より渡されたデバイス種別（通常デバイス、高速デバイス、論理ディスク装置のいずれか）から、転送元、あるいは、転送先のデバイスに論理ディスク装置を含んでいるか否かを判断する。論理ディスク装置を含む場合は、本実施例の場合、ボリュームマネージャ、及び、ディスクドライバに後述するバスアービタ１０５にインターフェースを設けていないので、バスアービタ１０５によるデータ転送機能は利用せず、データ転送ドライバ２０３が管理するシステムのメモリ１０２とデバイスドライバ（ディスクドライバ２０５、高速デバイスドライバ２０６）を利用して、ステップ１００４でデバイスを制御する。

【００９５】ステップ１００４に示す処理につき説明する。ところで、データ転送ドライバ２０３は、データ転送処理のためのスタテイックなメモリ領域をシステム上に確保している。このメモリはシステム初期化時に例えば2 MB 物理アドレス上連続に、ページング対象外ページとして確保されていて、カーネル空間として仮想空間にマップされている。これを2 面バッファとして運用してデータ転送を行う。ストライプ構成された論理ディスク装置のチャンクサイズが１２８ＫＢであれば8 チャンク分の入出力サイズの指定が可能である。この処理の特徴的なところは制御プログラム２０１より発せられた、１本のコンテキストで非同期のダブルバッファ処理を行うことである。論理ディスク装置はボリュームマネージャ２０４により管理され、非同期インターフェースを持つ複数のディスク装置の入出力を管理するが、すでに説明したように、論理ディスク装置の入出力結果を同期的に利用者に返さねばならぬため非同期のデバイス同士の同期を取っている。このため、データ転送ドライバ２０３が論理ディスク装置にアクセスするときはコールして、その処理の終了を待たねばならない。従って、２面バッファの非同期運用を行うため本処理は、以下のようにしている。

26

27　　28

[転送元が論理ディスクの場合]

論理ディスク装置からの第１バッファへの読み出し
↓
「高速デバイスへの第１バッファからの書き込み（非同期）
ル｜論理ディスク装置からの第２バッファへの読みし出し
プ｜高速デバイスへの第１バッファからの書き込みの待ち合わせ
｜高速デバイスへの第２バッファからの書き込み（非同期）
｜論理ディスクからの第１バッファへの読みだし
｜高速デバイスへの第２バッファからの書き込みの待ち合わせ
－

[転送元が高速デバイスの場合]

高速デバイスからの第１バッファへの読み出し（非同期）
↓
高速デバイスからの第１バッファへの読み出しの待ち合わせ
↓
「高速デバイスからの第２バッファへの読み出し（非同期）
ル｜論理ディスクへの第１バッファからの書き込み
プ｜高速デバイスからの第２バッファへの読み出しの待ち合わせ
｜高速デバイスからの第１バッファへの読みだし（非同期）
｜論理ディスクへの第２バッファからの書き込み
｜高速デバイスの第２バッファへの読みだしの待ち合わせ
－

【００９６】一方、データ転送対象デバイスに論理ディスク装置を含まない時、高速デバイス同士のデータ転送、あるいは、ディスクストライピングが有効でないケースのディスク装置と高速デバイスとのデータ転送がこれに当たるが、この場合は、データ転送機能を有するバスアービタ１０５を用いてデータ転送を行う。以下、この場合の動作について説明する。

【００９７】ステップ１００２において、データ転送ドライバ２０３によりデバイス間の転送速度の差がないと判断された場合、つまり、高速デバイス同士のデータ転

送の場合について説明する。高速デバイス１０８はＤＭＡコントローラを有するので、バスアービタ１０５によりハンドシェークすることでメモリ１０２を介在させずにデータ転送を行うことができる。ステップ１００５でデータ転送ドライバはシステムのメモリ空間にマップされたバスアービタ１０５のソースレジスタ４０１、デスティニレジスタ４０２に、転送元、転送先となる高速デバイス１０８のローカルバス１０７上のＩＤ（デバイスアドレス）を登録する。さらに、転送元の高速デバイス１０８に対し指定された入出力サイズ、あらかじめ、システムから与えられた未使用の物理メモリ空間のアドレスでリード要求を発行する。また、転送先の高速デバイス１０８に対し、同様に、サイズ、アドレスを渡し、ライト要求を発行する。これにより、２つのデバイスからバスにバス獲得要求がだされるが、該ＩＤはソースレジスタ４０１、デスティニレジスタ４０２に登録されている。さらに、ソースレジスタ４０１、デスティニレジスタ４０２にメモリ１０２の使用ビット（このビットについては後述する）は立っていないので、バスアービタ１０５は、ステップ１１０２、双方のバス要求が揃った時点で１１０４双方のバスグラント信号をアサートすることにより、メモリの介在なしのデータ転送が行われる。データ転送の終了の後、データ転送ドライバはアービタのソース／デスティニレジスタをクリアする。

【００９８】データ転送ドライバ２０３は、ステップ１００２でデバイス間の転送速度に差があると判断された場合、つまり、ディスク装置１１１と高速デバイス１０８の一方がが転送元で他方が転送先となっている場合には、システムが提供するメモリアロケータにより、物理アドレス上連続でページアウトされないメモリ１０２のエリアを、入出力サイズ分確保し、ステップ１００３でバスアービタ１０５のソースレジスタ４０１、デスティニレジスタ４０２に、各々のＩＤとメモリ使用ビットフラグをＯＲして登録する。この後、ステップ１００５で双方のデバイスに転送データサイズ、確保したメモリのアドレスを渡し、各々、リード／ライト要求を発行する。バスアービタ１０５は、ステップ１１０２でメモリ使用指示があるので、ソースデバイスからのバス要求に対しては、ステップ１１０３でバスグラント信号をアサートし、ソースデバイスのデータ転送終了を意味するビットをソースレジスタ４０１にＯＲする。転送先のデバイスは、バス要求が認められるまでバス要求を繰り返す。バスアービタ１０５は、バスフリーでかつソースデバイスのデータ転送終了が転送終了ビットがＯＮであることを確認すると、ステップ１１０６でグラント信号をアサートする。これにより、転送先デバイスは転送用に確保したメモリ１０２に転送されている転送元のデータを自デバイスに書き込む。これによりデータ転送が行われる。データ転送の終了の後、データ転送ドライバ２０３はバスアービタ１０５のソースレジスタ４０１、デスティニレジスタ４０２をクリアし、データ転送のために確保したメモリ１０２を解放する。

【００９９】以上のように、データ転送ドライバがデータ転送用の2面バッファを確保してデータ転送を行うようにしたので、入出力装置間のデータ転送を、より高速で行うことができる。即ち、システムのメモリ管理への負担の少ない、言い替えるとシステムのメモリ管理へ負担を掛けることによる性能低下を防ぐデータ転送方式を提供するすることができる。

【０１００】また、各入出力装置の制御をする制御装置に設けられたＤＭＡコントローラをコントロールしメモリを介さずに、各々の入出力装置のデータ転送をバス上で同期させるコントローラを主たる入出力バスに備えることにより、無駄なメモリへのデータ転送を除くことができる。

【０１０１】また、データ転送を行う制御プログラム（データ転送ドライバ）が、入出力装置間の速度差を吸収するために、主たる入出力バスに接続されるデータ転送コントローラ（バスアービタ）にシステムからメモリを確保して与えるようにしたので、データ転送コントローラはこのメモリを利用してデバイス間の転送速度調節をすることことができるようになり、入出力装置間のデータ転送を効率良く高速に行える。

【０１０２】実施例５．本実施例は、実施例1～実施例４で構成された入出力システムにおいて、各デバイスの持つＤＭＡコントローラを制御することにより（当然のことながら、デバイス間でのＤＭＡ転送レートは、一致させる必要はある）、メモリ１０２を介さずに各々のデバイスのデータ転送をバス上で同期させる制御ローカルバス１０７上に備えるようにして、無駄なメモリ１０２へのデータ転送を避ける例である。本実施例においては、各々のＳＣＳＩバス上の異なるディスク装置で4つのストライプディスクを持つ論理ディスク装置を構成することが多いが、データ整理のために論理ディスク装置の配置を変更する必要が発生することも多い。この時、ボリュームマネージャ２０４は、データ移動先の対応するディスク装置が同一のＳＣＳＩバス上にあれば、全てのストライプディスク装置に対して移動先のディスク装置を複写先として、ＳＣＳＩのコピーコマンドを発行することにより、データを複写し、複写元を未使用とすることにより、論理ディスク装置のデータ配置を変更することができる。

【０１０３】この実施例は、例えば、主たる入出力バス（ローカルバス）に複数のＳＣＳＩバスが接続され、各々のＳＣＳＩバスには複数のディスク装置が接続され、さらに、各々のＳＣＳＩバスから１つずつディスク装置を使用するように複数のストライプ構成の論理ディスク装置を構成したシステムに適用される。このようなシステムにおいて、論理ディスク装置間のデータ転送が必要となったときは、同一のバスに接続されたディスク装置

は、２つの論理ディスク装置は同一のストライプ幅で構成されているので、各々対応するストライプ同士で、ＳＣＳＩのコピーコマンドを利用することにより、この論理ディスク装置間でデータ転送を行うことにより、高速で、システムへの負担の少ないデータ転送を実現できる。また、高速デバイスをＳＣＳＩコントローラ機能を備えた複数の入出力ポートを有する高速デバイスとすることによって、種類の異なる入出力装置間にも適用できる。

【０１０４】なお、上記各実施例においては、入出力インタフェースとしてＳＣＳＩを例にとり説明したが、インタフェースは特にＳＣＳＩである必要はなく、他のインタフェースを用いてもよいことは言うまでもない。また、上記実施例5は、ディスク装置間の複写をＳＣＳＩのＣＯＰＹコマンドを用いて説明したが、ディスク制御装置にＳＣＳＩのＣＯＰＹコマンド相当の機能を実現すればよい。

【０１０５】

【発明の効果】以上のように本発明によれば、システムに接続されているディスク装置の性能特性を性能収集手段がシステム管理者の与えるデータもしくは性能計測手段により収集し、収集したデータを基に、論理ディスク構成手段により論理ディスクを構成し、論理ディスク制御手段により論理ディスクを動作させるので、論理ディスクを構成する各ディスクの1ストライプ分のデータ入出力に要する応答時間が均等となり、論理ディスク装置の性能向上を図ることができる。

【０１０６】また、入出力命令の処理に要する入出力装置の実行時間を制限時間設定手段により規定し、制限時間が経過すると入出力命令を終了させる手段を設けたので、入出力動作の同期を時間を基準にコントロールすることが可能となり、アプリケーションプログラムでの余分な待ち時間を削減することができる、且つ大量のデータ転送を行う際の処理を効率良く行うことができる。

【０１０７】また、入出力制御装置に入出力命令の制限時間を設定するタイマとこのタイマが設定時間を経過すると実行中の入出力命令を終了させる手段を設けるようにしたので、処理の実現を容易に行うことができる。

【０１０８】また、複数のディスク装置で構成される論理ディスク装置において、この複数のディスク装置を並列動作させる場合に、ディスク装置間の入出力完了の同期を入出力動作の制限時間を設定することによりとるようにすることにより、ユーザの希望した時間までに、この論理ディスク装置に対する入出力処理を完了することができる。

【０１０９】さらに、ストライプ構成された論理ディスク装置に、複数のディスク装置間の同期方法として、この実施例で説明した同期方法を適用することにより、設定する制限時間により論理ディスク装置に対する入出力動作の応答時間は保証され、かつ、入出力動作の完了ステータスのばらつきの少ない、つまり、応答時間に見合ったデータ量設定であれば、全てのディスク装置における入出力動作が完了させることができる。

【０１１０】また、性能計測手段でディスク装置の性能を計測する際に、入出力命令に対する応答時間が最短あるいは最長となる条件を設定して行うようにしたので、論理ディスク構成装置が論理ディスクを構成する際に性能が最適になる論理ディスク装置を構成することができる。

【０１１１】また、システムにディスク装置を導入する際に、初期化に一部として性能計測手段により性能を計測するようにしたのでシステムとしての効率を向上させることができる。

【０１１２】また、性能計測手段に入出力装置の性能だけでなく、入出力バスの性能を計測する手段を設け、性能収集手段を構成するようにしたのでシステム全体の入出力性能を向上させることができる。

【０１１３】また、論理ディスク構成手段が、1台のディスク装置に対する1回の入出力命令で行うデータ量をそのディスク装置の1トラックに収まるように論理ディスク構成データを構成するようにしたので、論理ディスク装置を構成するディスク装置のシーク方向が順方向と逆方向での性能差がなくなる。

【０１１４】また、論理ディスク構成手段が内外周での性能に差があるディスク装置が論理ディスク装置に含まれているときは、内外周を交互に組み合わせて論理ディスクを構成するようにするので、各ストライプに対して均等な処理時間となり、処理効率を向上させることができる。

【０１１５】また、論理ディスク制御手段がディスク装置に指令を出すときに、各ディスク装置のヘッド位置を考慮して入出力実行時間が最少となるように制御するので処理効率が向上する。

【０１１６】また、同一のディスク装置に対する複数の入出力命令を纏めるようにしたので、命令実行に伴う処理時間を減少させることができるので入出力処理を向上させることができる。

【０１１７】また、論理ディスク装置の構成を自動的に行うようにしたので、論理ディスクの構成が容易になる。

【０１１８】また、データ転送ドライバが入出力命令の対象賭する入出力装置に論理ディスク装置が含まれている場合は、２面バッファをシステムメモリ上に確保してデータ転送を行うようにしたので、アプリケーションプログラムでメモリを確保する必要がなく、通常アプリケーション上のバッファメモリを介して行われていたデバイス間のデータ転送を、システムの仮想記憶に負担を掛けないページ管理外のメモリを使用して、あるいは、若干の特殊ハードウェアを付加することによりメモリを使用せずにデータ転送することを可能にし、大量高速デー

33

タ転送をシステムへの負担を軽減し、なおかつ、高速に実行できる。

【０１１９】また、入出力バスの制御をするバスアービタに入出力装置の並列動作をさせる手段を設け、入出力命令の対象に論理ディスク装置が含まれていない場合には、このバスアービタを動作させるようにしてシステムメモリを使用せずにデータ転送をすることができるようにしたので、データ転送を高速にできる。

【０１２０】また、データ転送ドライバに入出力命令の対象となる入出力装置間に速度性能差がある場合には、システムメモリにデータ転送用のバッファを確保してバスアービタを動作させるようにしたので、アプリケーションでメモリを確保する必要のない効率のよいデータ転送を行うことができる。

【０１２１】また、ディスク装置を制御するディスク制御装置に同一バスに接続されたディスク装置間の複写手段を設けるようにしたので、論理ディスク装置のバックアップが容易に実現できる。

【図面の簡単な説明】

【図1】　本発明による入出力処理システムの適用されるシステムのハードウエアの構成の一例を示す図である。

【図2】　本発明による入出力処理システムの適用されるシステムのソフトウェアの構成の一例を示す図である。

【図3】　本発明による性能計測手段によって作成されたディスクの性能データの一例を示す図である。

【図4】　本発明によるボリュームマネージャの行う論理ディスク装置を自動的に構成するための処理の流れを

34

示す図である。

【図5】　本発明による論理ディスク管理データの一例を示す図である。

【図6】　本発明による論理ディスク構成データの一例を示す図である。

【図7】　本発明におけるよるディスクコントローラの構成を示すブロック図である。

【図8】　本発明における論理ディスク装置の動作状況の一例を示す図である。

【図9】　本発明におけるバスアービタの構成を示すブロック図である。

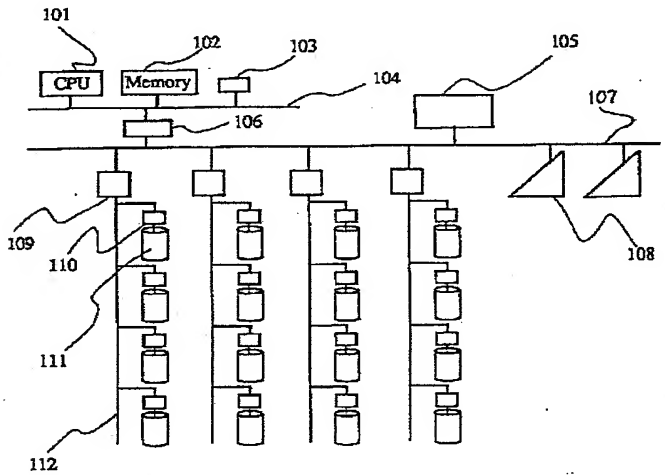
【図10】　本発明によるデータ転送ドライバの処理の流れの一例を示す図である。

【図11】　本発明におけるバスアービタの処理の流れの一例を示す図である。

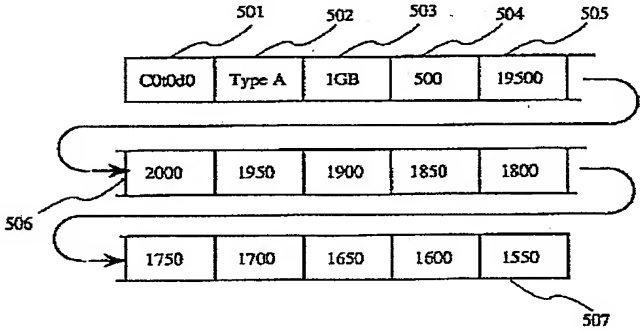
【符号の説明】

１０５　バスアービタ、１０８　高速デバイス、１０９　ＳＣＳＩホストアダプタ、１１０　ディスクコントローラ、２０１　データ転送制御プログラム、２０３　データ転送ドライバ、２０４　ボリュームマネージャ、２０８　性能計測手段、２０９　性能データ、２１０　ディスク管理データ、２１１　論理ディスク構成データ、２１２　論理ディスク入出力インターフェース、２１３　論理ディスク構成手段、３０２　ＳＣＳＩコントロールロジック、３０４　ディスクコントロールロジック、３０５　タイマ、４０１　ソースレジスタ、４０２　デスティニレジスタ、４０３　バスアービトレーションロジック。

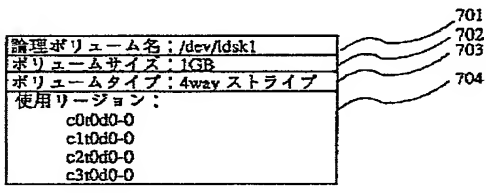
【図1】



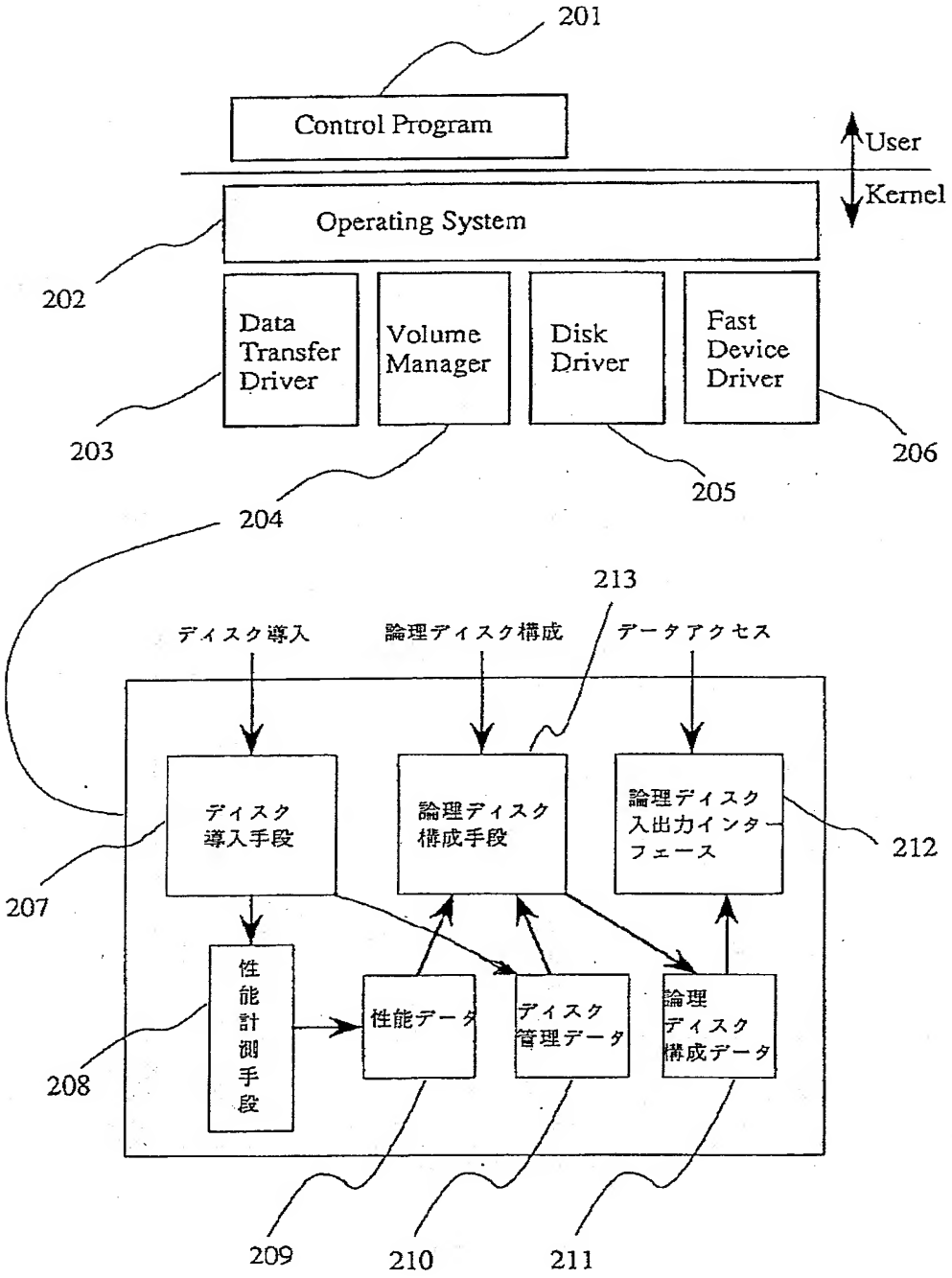
【図3】



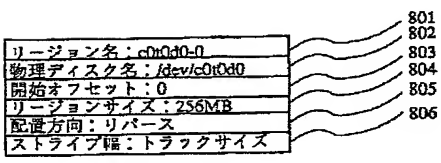
【図5】



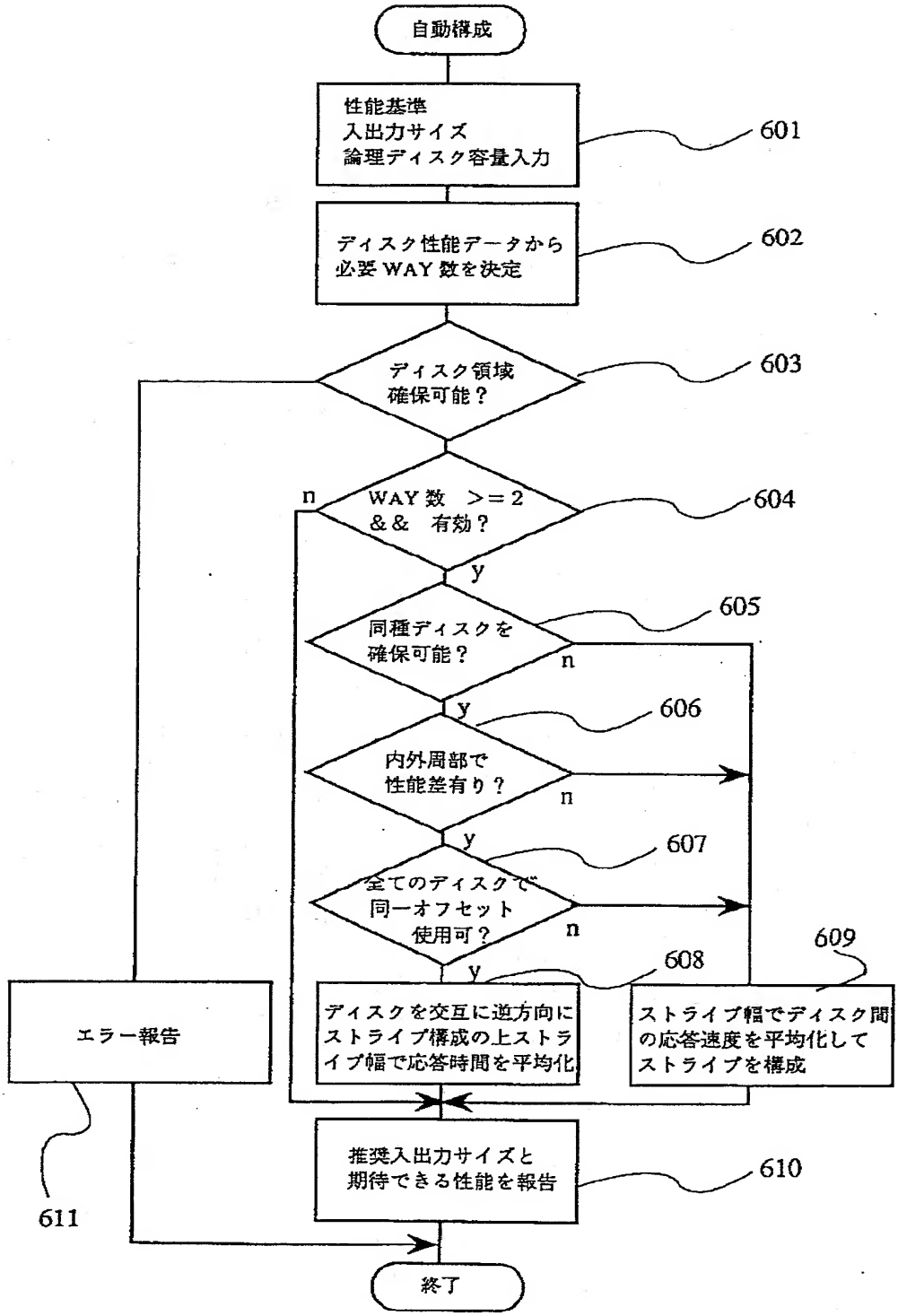
【 図2 】


201
Control Program
User
Kernel
Operating System
202
Data Transfer Driver
Volume Manager
Disk Driver
Fast Device Driver
203
204
205
206
213
ディスク導入
論理ディスク構成
データアクセス
ディスク導入手段
論理ディスク構成手段
論理ディスク入出力インターフェース
207
212
性能計測手段
性能データ
ディスク管理データ
論理ディスク構成データ
208
209
210
211

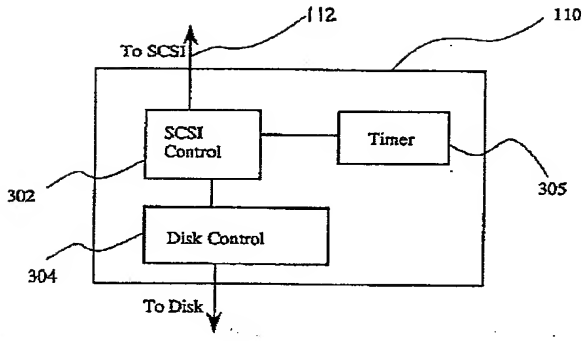
【 図6 】


リージョン名：c0t0d0-0
物理ディスク名：/dev/c0t0d0
開始オフセット：0
リージョンサイズ：256MB
配置方向：リバース
ストライプ幅：トラックサイズ
801
802
803
804
805
806

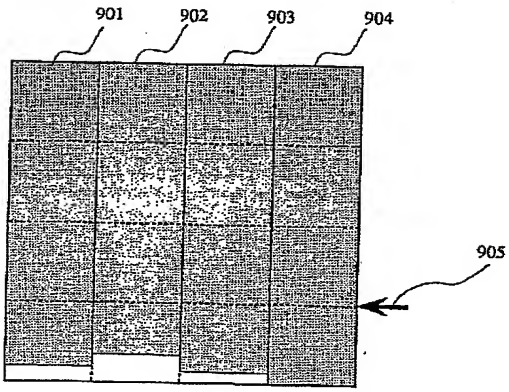
【 図4 】


自動構成
性能基準
入出力サイズ
論理ディスク容量入力
601
ディスク性能データから
必要 WAY 数を決定
602
ディスク領域
確保可能？
603
n
WAY 数　＞＝2
＆＆　有効？
604
y
同種ディスクを
確保可能？
605
n
y
内外周部で
性能差有り？
606
n
y
全てのディスクで
同一オフセット
使用可？
607
n
y
608
609
エラー報告
ディスクを交互に逆方向に
ストライプ構成の上ストラ
イプ幅で応答時間を平均化
ストライプ幅でディスク間
の応答速度を平均化して
ストライプを構成
推奨入出力サイズと
期待できる性能を報告
610
611
終了

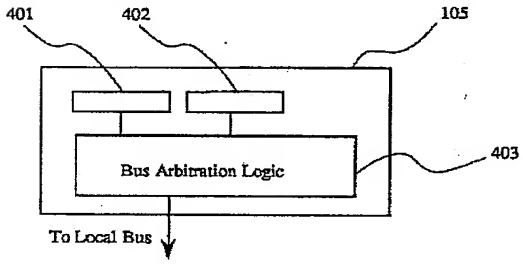
【 図7 】


112
110
To SCSI
SCSI Control
Timer
305
302
Disk Control
304
To Disk

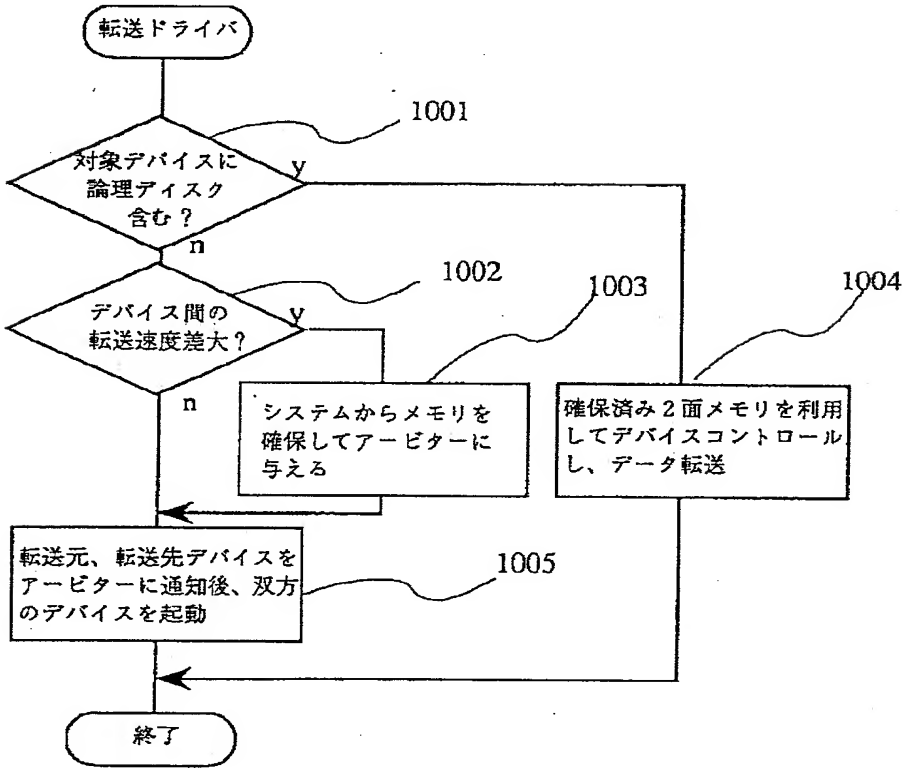
【 図8 】


901
902
903
904
905

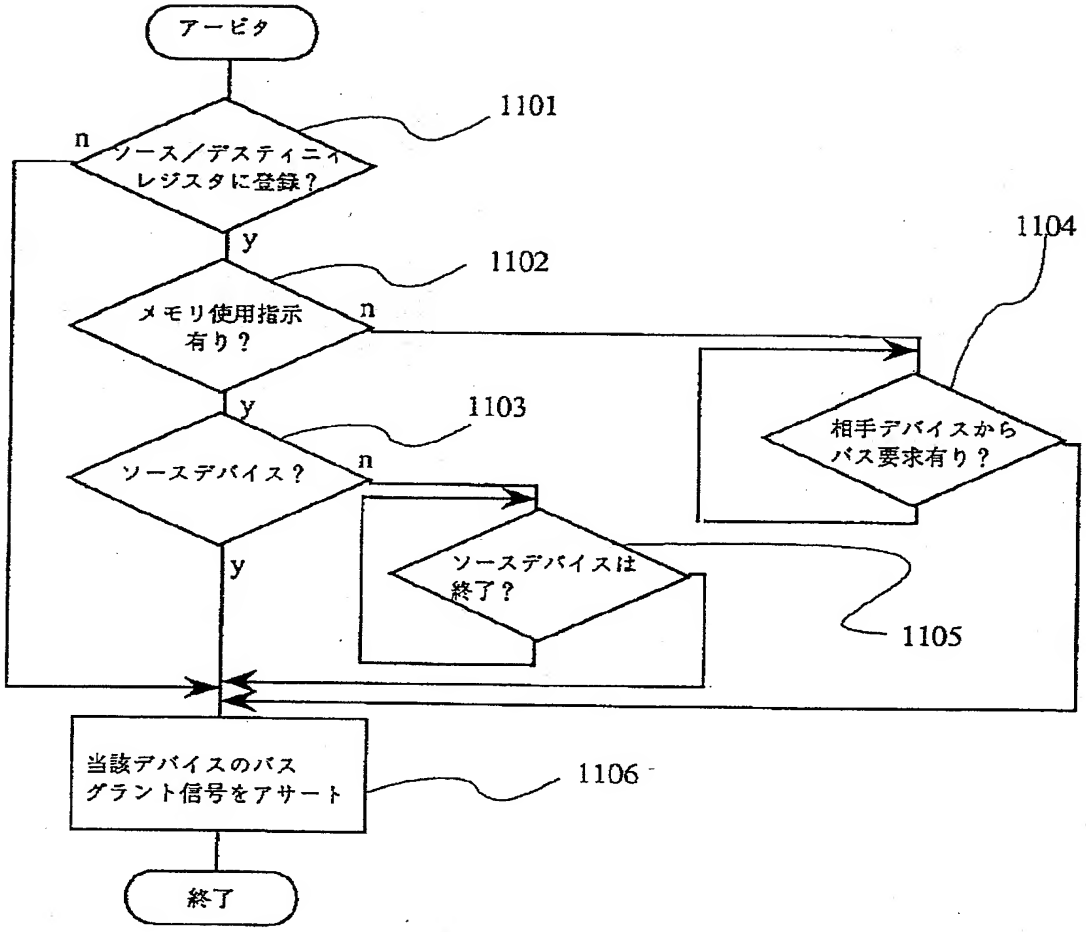
【 図9 】


401
402
105
Bus Arbitration Logic
403
To Local Bus

【 図10 】


転送ドライバ
1001
対象デバイスに論理ディスク含む？
y
n
1002
デバイス間の転送速度差大？
y
n
1003
1004
システムからメモリを確保してアービターに与える
確保済み２面メモリを利用してデバイスコントロールし、データ転送
転送元、転送先デバイスをアービターに通知後、双方のデバイスを起動
1005
終了

【 図1 1 】


アービタ
1101
ソース／デスティニィ
レジスタに登録？
n
y
1102
メモリ使用指示
有り？
n
1104
y
1103
ソースデバイス？
n
相手デバイスから
バス要求有り？
y
ソースデバイスは
終了？
1105
当該デバイスのバス
グラント信号をアサート
1106
終了